COMANDI STATISTICA

❖ **Statistica descrittiva**

NOME	**COMANDO**
Cancella tutto	rm(list=ls())
Vettore	vettore<-c()
Lunghezza	str() → num [1:10] length() → 10
Dati in ordine crescente	sort()
Frequenza assoluta (moda)	table(a)
Frequenza relativa	prop.table(b)
Frequenza percentuale	prop.table(b)*100
Media, xbar	mean()
Mediana	median()
Media ponderata, wa	sum(a*freq_a)/tot_a) con tot_a = sum(freq_a)
Deviazione standard, s	sd o sqrt(var)
Varianza, s^2	var()
Coefficiente di variazione, cv	sd()/mean() con mean in valore assoluto
Boxplot	boxplot(a,horizontal=TRUE,col="yellow") boxplot(a,b,horizontal=TRUE, main="titolo", names=c("a","b"), col=c("orange","lightblue"))
Scatterplot → length uguale	plot(a,b,col="red")
Confronto tra percentili → length non uguale	qqplot(a,b,col="red")
Istogramma	hist() breaks=numero, è il n°di classi
Diagramma a torta	pie(tabella frequenze)
Diagramma a barre	barplot(tabella frequenze)
Min, 1st qu., Mediam, Mean, 3rd qu., Max	summary()
Percentile	quantile(vettore,k%)
Covarianza	cov(x,y)
Coefficiente di corr di Pearson, r	cor(x,y) devo avere sd() e cov()
Retta di regressione	lm(y~x) con y=output e x=input y<-function(x)(m*x+q)
Sovrapporre	abline()
Valori attesi	attesi<-round(predict(reg), digits=0) setNames(attesi, vettore richiesto)
Per agire con un comando sulla singola variabile del dataframe	comando(dataframe $ singola variabile)
Estremi del range	range(dataframe $ singola variabile)
Più piccolo e più grande valore di tutta la tabella	range(dataframe)
Ampiezza del range	range(dataframe $ singola variabile)[2]-range(dataframe $ singola variabile)[1]
Intervallo interquartile (distanza tra q_3-q_1)	IQR(dataframe $ singola variabile)
Parametri di centralità per tutte le coppie di variabili	summary(dataframe)
Parametri di dispersione per tutte le coppie di variabili (covarianza e coeff. Di Pearson)	cov(dataframe) cor(dataframe)

Gavioli Alice

❖ Probabilità

P(A)	k/n
Unione	$A \cup B = \{x \in \Omega;\ x \in A \text{ o } x \in B\}$
Intersezione	$A \cap B = \{x \in \Omega;\ x \in A \text{ e } x \in B\}$
Complementare	$A^c = \{x \in \Omega;\ x \notin A\}$ *$(A \cup A^c) = \Omega$* *$(A \cap A^c) = 0$*
Assiomi di Kolmogorov Spazio di probabilità (Ω, a, P)	• $0 \leq P(A) \leq 1$ • $P(\Omega) = 1$ • $A \cap B = 0 \rightarrow P(A \cup B) = P(A) + P(B)$
1. Probabilità del complementare	$P(A^c) = 1 - P(A)$
2. Probabilità dell'evento impossibile	$P(0) = 0$
3. Probabilità dell'evento certo	$P(B) = P(B \cap A) + P(B \cap A^C)$
4. Probabilità di ordinamento	$A \subset B \rightarrow P(A) \leq P(B)$
5. Probabilità dell'unione di eventi non disgiunti	$P(A \cup B) = P(A) + P(B) - P(A \cap B)$
5bis. Estensione punto 5	$P(A \cup B \cup C) = P(A) + P(B) + P(C)$ $- P(A \cap B) - P(A \cap C) - (B \cap C) + P(A \cap B \cap C)$
Eventi indipendenti	$P(A \cap B) = P(A) * P(B)$ solo se A e B sono indipendenti
Spazio di probabilità uniforme	$P(A) = P(\|A\|) * p = \|A\| / \|\Omega\|$
	$P(A \cap B) \leq P(A) \leq P(A \cup B)$
Valore atteso=media, mu, E[x]	sum(x*x) n*p
Varianza, sigma2	sum(x-mu)^2 x n*p(1-p)
Deviazione standard, sigma	sqrt(sigma2)
n!	factorial(n)
(n k)	choose(n,k)

P(x=k)	$p(k) = P(x=k) = (n\ k) p^k (1-p)^{n-k}$ dbinom(k,size=n,prop=p)
$P(a \leq x \leq b)$	sum(dbinom(a:b,size=n,prop=p))
$P(a \leq x < b)$	sum(dbinom(a:b-1,size=n,prop=p))
$P(a < x \leq b)$	sum(dbinom(a+1:b), size=n, prob=p))
$P(x \geq a)$ o $1-P(x=0)$	sum(dbinom(a:n,size=n,prop=p))
$P(x > a)$	sum(dbinom(a+1:n,size=n,prop=p))

$P(z > a)$ o $1-P(z \leq a)$	1-pnorm(a,mean=mu,sd=sigma) pnorm(a,mean=mu,sd=sigma,lower.tail=FALSE)
$P(a < Z < b)$	pnorm(b,mean=mu,sd=sigma)- pnorm(a,mean=mu,sd=sigma)
$P(x < a)$	pnorm(a,mean=mu,sd=sigma)

DISTRIB.	**DENSITA' pdf**	**RIPARTIZIONE cdf**	**QUANTILI ORDINE α**
x~unif[a,b] **uniforme**	dunif(x,min=a,max=b)	punif(x,min=a,max=b)	qunif(α,min=a,max=b)
x~N(μ,δ^2) **normale standard**	dnorm(x,mean=μ,sd=δ)	pnorm(x,mean=μ,sd=δ)	qnorm(α,mean=μ,sd=δ)
x=B(n,p) **quantili**	dbinom(x,size=n,prob=p)	pbinom(x,size=n,prob=p)	qbinom(α,size=n,prob=p)

❖ **Statistica inferenziale**

Media μ e deviazione standard σ NOTE

Valore atteso	E[x]= μ
Varianza	Var(x)= σ^2/n
Teorema del limite centrale	x~N(μ,σ^2/n) con n ≥ 30 quindi non necessaria normalità del campione
Errore statistico (E>0)	E = Z* σ /√n con Z*=$Z_{1-\frac{\alpha}{2}}$ qnorm(1-(alpha/2),mean=0,sd=1)
Livello di fiducia	1-α=% → α=1-cl
Intervallo di confidenza, cl	μ = xbar ± E IC=(xbar-E,xbar+E) IC<-xbar+c(-E,E)

Media μ e varianza σ^2 NON NOTE, distr norm

Valore atteso	E[x]= 0
Varianza	Var(x) = n/n-2
t di student	t_{n-1}
Errore statistico	E = t* s/√n con t*=$t_{1-\frac{\alpha}{2},n-1}$ qt(1-alpha/2, df=n-1) se n>30 $t_{1-\frac{\alpha}{2},n-1}$ sostituito con $Z_{1-\frac{\alpha}{2}}$
Livello di fiducia	1-α/2 con n-1 gradi di libertà
Intervallo di confidenza, cl	μ = xbar ± E IC=(xbar-E,xbar+E) IC<-xbar+c(-E,E)

Proporzione p di una popolazione bernoulliana

Proporzione campionaria, phat	phat=n°successi nel campione / n
Ipotesi di lavoro	n*phat ≥ 5 e n*(1-phat) ≥ 5
Errore statistico	E = $Z_{1-\frac{\alpha}{2}}$ * $\sqrt{\frac{phat*(1-phat)}{n}}$ qnorm(1-(alpha/2),mean=0, sd=1) ↔ qnorm(1-(alpha/2))
Livello di fiducia	1-α=% → α=1-cl
Intervallo di confidenza, cl	μ = phat ± E IC=(phat-E,phat+E) IC<-phat+c(-E,E)

Stima della varianza, distr norm e varianza nota

Distribuzione chi-quadro con n gradi di libertà	dchisq(x,df=n)
Quantili	l*=X^2 α/2, n qchisq(alpha/2,df=n) r*=X^2 1-α/2, n qchisq(1-alpha/2,df=n)

Stima della varianza, distr norm e varianza NON nota

Livello di fiducia	1-α
Quantili	l*=X^2 α/2, n-1 qchisq(alpha/2, df=n-1) r*=X^2 1-α/2, n-1 qchisq(1-alpha/2, df=n-1) qchisq(alpha/2, df=n-1, lower.tail = FALSE)
Intervallo di confidenza	$(\frac{(n-1)s^2}{r*};\frac{(n-1)s^2}{l*})$ IC<-(n-1)*s2*c(1/rstar, 1/lstar)

Livelli di confidenza minori forniscono stime intervallari più precise

TEST di ipotesi con popolazione bernoulliana $Z = \frac{p-p0}{\sqrt{\frac{p0(1-p0)}{n}}}$

Test a una cosa destra prop.test(x,n,p=p0,alternative="**greater**") con x=n° successi del campione, round(n*phat) n=ampiezza del campione p0=parametro teorico di confronto	H0:p=p0 HA:p>p0 Pvalue=P(Z>z)
Test a una coda sinistra prop.test(x,n,p=p0,alternative="**less**")	H0:p=p0 HA:p<p0 Pvalue=P(Z<z)
Test a 2 code prop.test(x,n,p=p0,alternative="**two.sided**")	H0:p=p0 HA:p≠p0 Pvalue=2P(Z>|z|)=2(1-P(Z<|z|))

TEST di ipotesi per media con varianza NOTA $Z = \frac{xbar-\mu0}{\frac{\sigma}{\sqrt{n}}}$

Test a una cosa destra ➢ pvalue<-1-pnorm(z) ➢ pvalue<-1-pnorm(z,mean=0,sd=1,lower.tail=FALSE)	H0:μ=μ0 HA: μ>μ0 Pvalue=P(Z>z)
Test a una coda sinistra pvalue<-pnorm(z,mean=0,sd=1)	H0:μ=μ0 HA: μ<μ0 Pvalue=P(Z<z)
Test a 2 code ➢ pvalue<-2*(1-pnorm(abs(z))) ➢ pvalue<-2*pt(abs(z),lower.tail=FALSE)	H0:μ=μ0 HA: μ ≠ μ0 Pvalue=2P(Z>|z|)=2(1-P(Z<|z|))

TEST di ipotesi per media con varianza NON NOTA $t = \frac{xbar-\mu0}{\frac{s}{\sqrt{n}}}$

Test a una cosa destra t.test(x,mu=mu0,alternative="**greater**")	H0:μ=μ0 HA: μ>μ0 Pvalue=P(T>t)
Test a una coda sinistra ➢ t.test(x,mu=mu0,alternative="**less**") ➢ pvalue<-pt(t,df=(n-1))	H0:μ=μ0 HA: μ<μ0 Pvalue=P(T<t)
Test a 2 code ➢ t.test(x,mu=mu0,alternative="**two.sided**") ➢ pvalue<-2*pt(abs(t),df=(n-1),lower.tail=FALSE) ➢ pvalue<-2*(1-pt(abs(t),df=(n-1)))	H0:μ=μ0 HA: μ ≠ μ0 Pvalue=2P(T>|t|)=2(1-P(T<|t|))

TEST di ipotesi per la varianza $(chi2) x^2 = \frac{(n-1)s^2}{\sigma0^2}$

Test a una cosa destra ➢ pvalue<-1-pchisq(chi2,df=n-1) ➢ pvalue<-pchisq(chi2,df=n-1, lower.tail=FALSE)	H0:$\sigma^2 = \sigma0^2$ HA:$\sigma^2 > \sigma0^2$ Pvalue=P(X>x^2)
Test a una coda sinistra pvalue<-pchisq(chi2,df=n-1)	H0:$\sigma^2 = \sigma0^2$ HA:$\sigma^2 < \sigma0^2$ Pvalue=P(X<x^2)
Test a 2 code pvalue<-2*min(pchisq(chi,df=(n-1),lower.tail=TRUE), pchisq(chi, df=(n-1),lower.tail=FALSE))	H0:$\sigma^2 = \sigma0^2$ HA:$\sigma^2 \neq \sigma0^2$ Pvalue=2min(P(X<x^2);P(X>x^2))

Test di ipotesi per la mediana, distrib NON nota = TEST DI WILCOXON

Test a una cosa destra wilcox.test(x,mu=m,alternative="**greater**")	H0:mediana=m HA:mediana>m
Test a una coda sinistra wilcox.test(x,mu=m,alternative="**less**")	H0:mediana=m HA:mediana<m
Test a 2 code wilcox.test(x,mu=m,alternative="**two.sided**")	H0:mediana=m HA:mediana≠m

❖ Test di confronto tra 2 proporzioni di 2 popolazioni bernoulliane

$$z = \frac{p1 - p2}{\sqrt{\frac{p(1-p)}{n1 + n2}}}$$

```
z test
```

❖ Test di confronto tra medie di 2 popolazioni

-varianza note

$$z = \frac{xbar - ybar}{\sqrt{\frac{\sigma x^2 + \sigma y^2}{nx + ny}}}$$

```
z test, con pvalue<-pnorm(z)
```

-varianze NON note, ma uguali

$$t = \frac{xbar - ybar}{sp\sqrt{\frac{1}{nx + ny}}} \quad sp = deviazione\ standard\ pooled$$

```
t.test(x,y,alternative="grater",var.equal=TRUE)
                        "less",
                        "two.sided",
```

-varianze NON note, NON uguali

$$t = \frac{xbar - ybar}{\sqrt{\frac{sx^2 + sy^2}{nx + ny}}} \quad f\ ha\ distribuzione\ di\ Fischer\ co\ nA - 1\ e\ nB - 1\ gradi\ di\ libert\text{à}$$

```
t.test(x,y,alternative="grater",var.equal=FALSE)
                        "less",
                        "two.sided",
```

❖ Test di confronto sulle varianze

$$f = \frac{s^2 A}{s^2 B}$$

```
var.test(A,B,alternative="grater")
                         "less")
                         "two.sided")
```

❖ Test di confronto tra mediane

```
Wilcox.test(x,y,alternative="grater")
                             "less")
                             "two.sided")
```

TEST di indipendenza n=(n°righe-1)*(n°colonne-1)	H0:le variabili sono indipendenti HA:le variabili non sono indipendenti costruiamo la tabella di contingenza: x<-rbind(rigasopra,rigasotto) chisq.test(x)
TEST di Shapiro-Wilk	H0:X ha distribuzione normale HA:X NON ha distribuzione normale shapiro.test(x)
TEST di adattamento	H0:X=X0 (modello compatibile con i dati) HA:X≠X0 (modello NON compatibile con i dati) chisq.test(z,p=pt) con z=frequenze,pt=elementi
TEST di Kolmogorov-Smirnov	H0:la distribuzione si adatta al modello HA:la distribuzione NON si adatta al modello ks.test(x,"nome cdf",distribuzione)
Confronto tra distribuzioni di 2 popolazioni	H0:X=Y (distrib. 2 popolaz. sono uguali) HA:X≠Y (distrib. 2 popolaz. NON sono uguali) ks.test(x,y)
TEST ANOVA a una via distrib. normale e stesse varianze (anche non note)	H0: $\mu=\mu 0=\cdots=\mu k$ HA:almeno una delle μ diversa dalle altre costruiamo un dataframe: l<-list(maggio=may,settembre=sep,dicembre=dec) d<-stack(l) oneway.test(colonna1~colonna2,data=d, var.equal=TRUE) var.equal=FALSE) → se non è ipotizzabile l'uguaglianza tra le varianze
TEST di Kruskall Wallis	H0: $\mu=\mu 0=\cdots=\mu k$ HA:almeno una delle μ diversa dalle altre costruiamo un dataframe: l<-list(maggio=may,settembre=sep,dicembre=dec) d<-stack(l) kruskal.test(colonna1~colonna2,data=d)
TEST di ipotesi per (rho)ρ=0 Pvalue piccolo → |t|grande → r prossimo a 1 Pvalue grande → |t|piccolo → r prossimo a 0	H0:ρ=0 (NON c'è correlazione lineare) HA:ρ≠0 (c'è correlazione lineare) Cor(x,y) $t=r\left(\sqrt{\frac{n-2}{1-r^2}}\right)$ $n-2$ *gradi di libertà* Pvalue=2P(T>|t|) → cor.test(x,y)
TEST di pendenza e intercetta = 0	H0: beta0=0 intercetta (beta1=0 pendenza) HA: beta0≠0 intercetta (beta1≠0 pendenza) reg<- lm(y~x) summary(reg)→ summary(lm(y~x))

Gavioli Alice